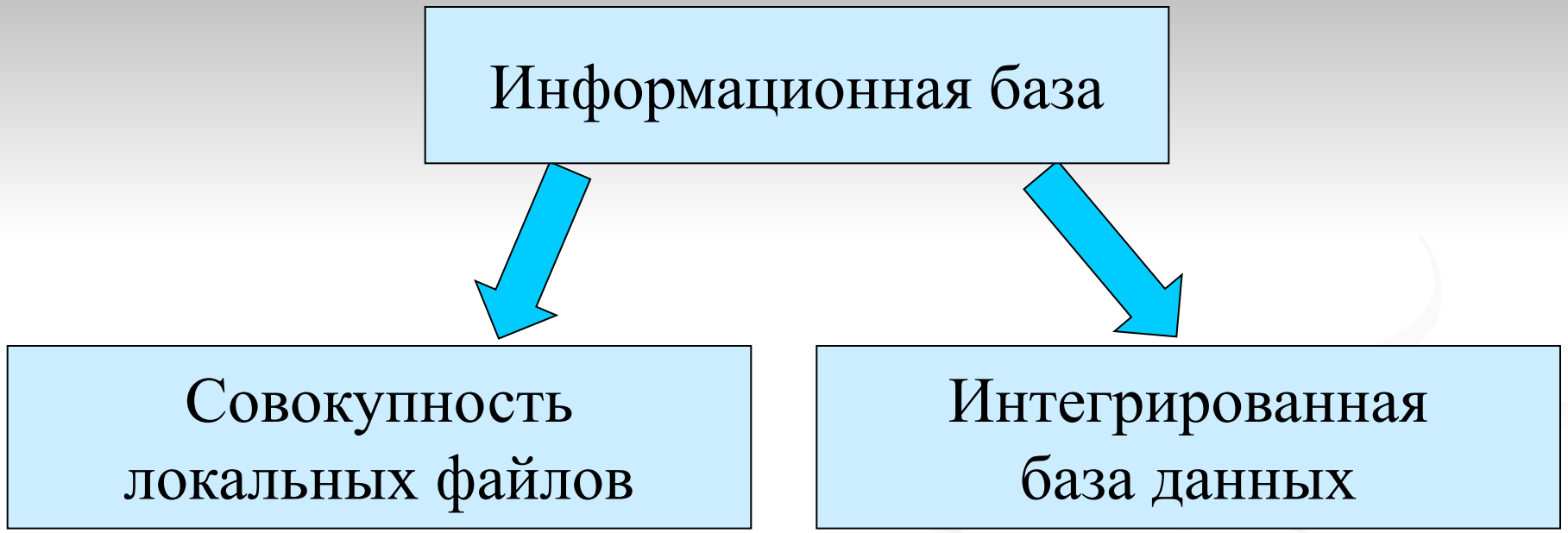
ТЕМА 8. Информационное обеспечение ИС.

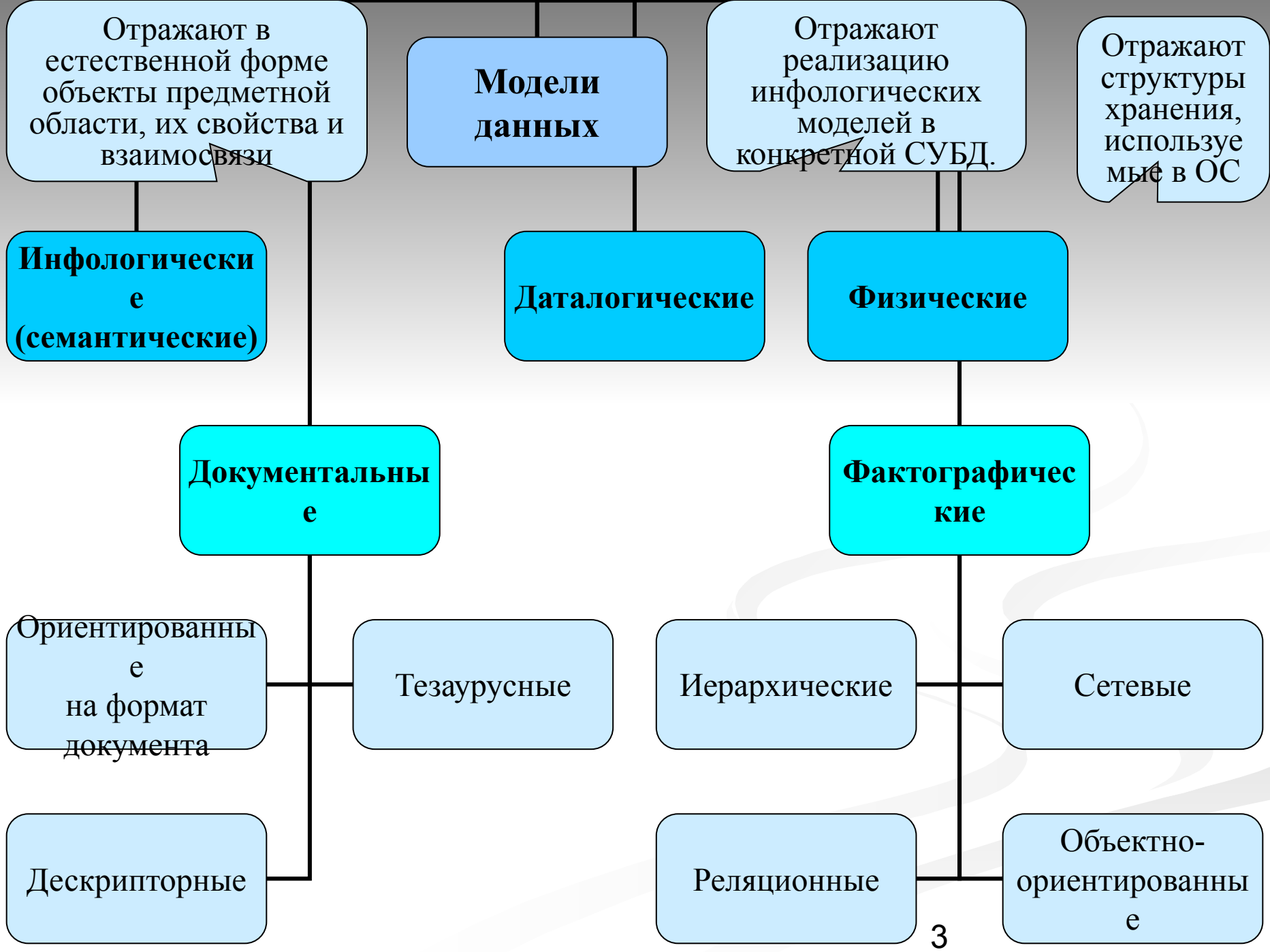
Лекция 20.
Внутримашинное ИО.
Информационные хранилища.

Способы организации информационной базы



Поддерживается функциональными пакетами прикладных программ

Основывается на использовании универсальных программных средств загрузки, хранения, поиска и ведения данных (СУБД).


Отражают в естественной форме объекты предметной области, их свойства и взаимосвязи
Модели данных
Отражают реализацию инфологических моделей в конкретной СУБД.
Отражают структуры хранения, используемые в ОС
Инфологические (семантические)
Даталогические
Физические
Документальные
Фактографические
Ориентированные на формат документа
Тезаурусные
Иерархические
Сетевые
Дескрипторные
Реляционные
Объектно-ориентированные

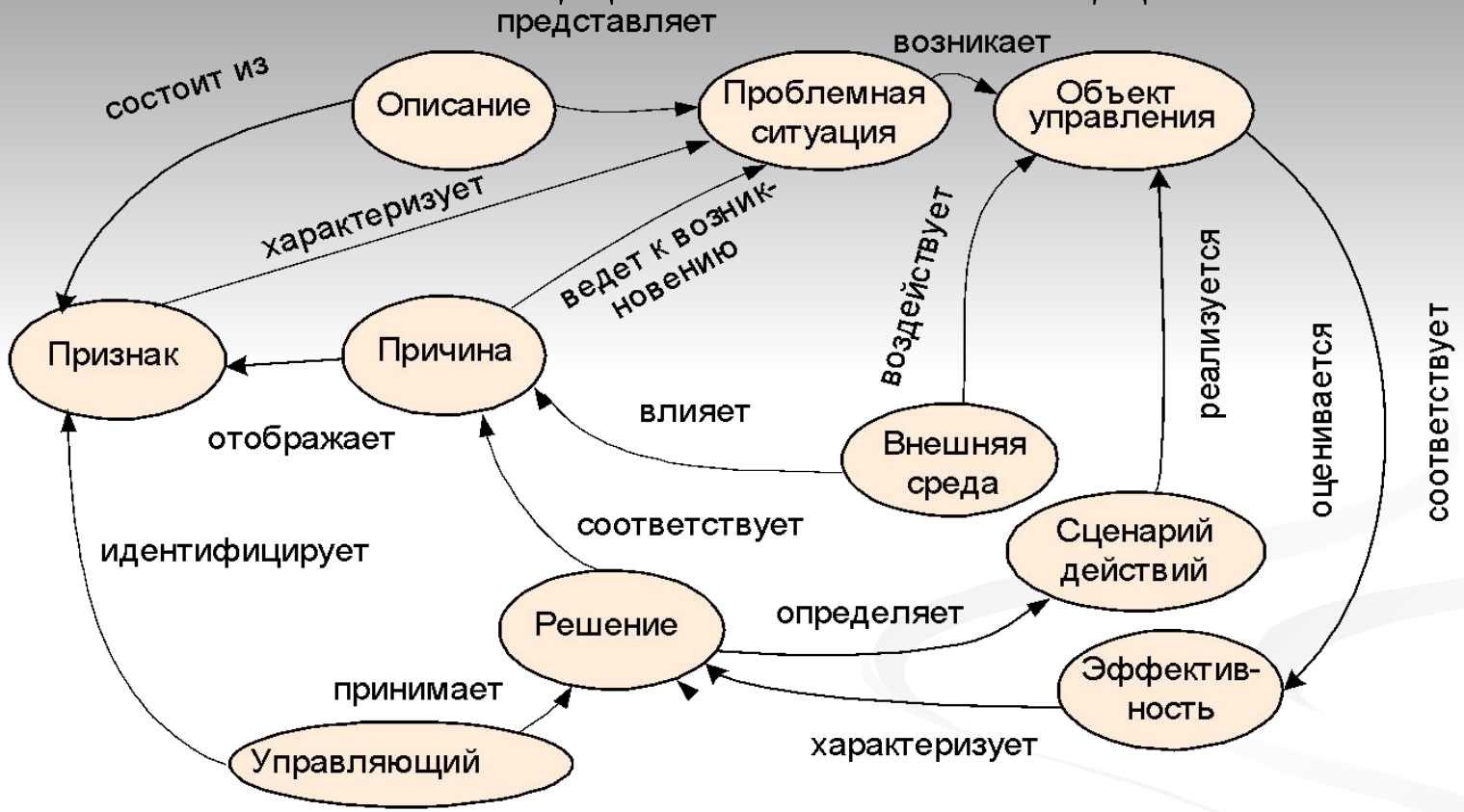
Классификация моделей данных

- ***Инфологические модели*** отражают в естественной форме объекты предметной области, их свойства и их взаимосвязи.
- ***Даталогические модели*** отражают реализацию инфологических моделей в конкретной СУБД.
- ***Физические модели*** оперируют категориями, касающимися организации внешней памяти и структур хранения, используемых в операционной системе.
- ***Документальные модели*** соответствуют представлению о слабоструктурированной информации, ориентированной в основном на свободные форматы документов, текстов на естественном языке.
  - ***Модели, ориентированные на формат документа***, связаны с использованием языков разметки (SGML, HTML, XML).
  - ***Тезаурусные модели*** основаны на принципе организации словарей, содержат определенные языковые конструкции и принципы их взаимодействия в заданной грамматике (системы-переводчики).
  - ***Дескрипторные модели*** основаны на использовании описателей (дескрипторов) документов. Дескриптор имеет жесткую структуру и описывает документ в соответствии с теми характеристиками, которые требуются для работы с документами в разрабатываемой БД.
- ***Фактографические модели*** отражают совокупность фактов – сведений о предметной области без привязки к документам.

Иерархическая модель базы данных

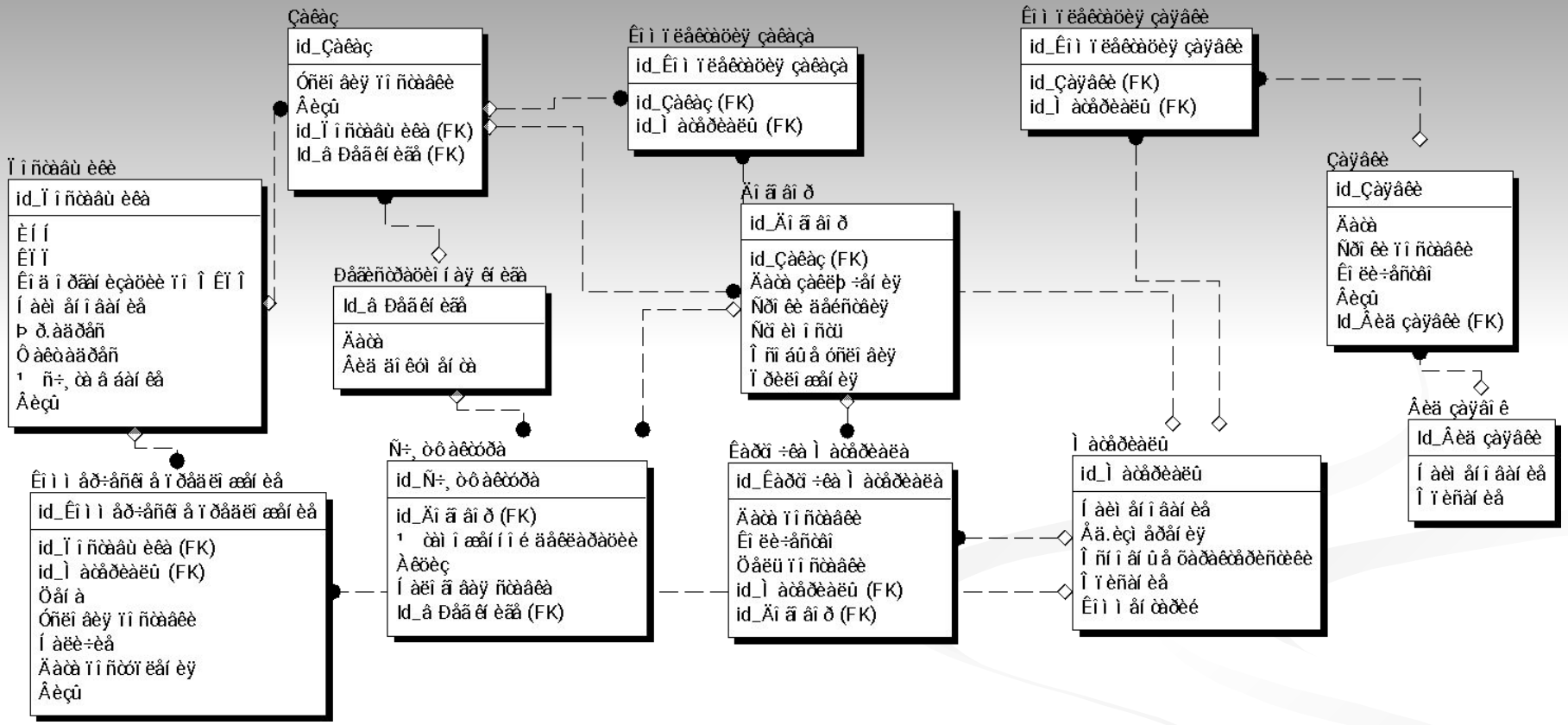
- **Достоинство:** экономичное использование ресурсов памяти и высокое быстродействие системы.
- **Недостаток:** жесткие связи и необходимость перепрограммирования базы данных при изменении модели.

Сетевая модель базы данных



- **Достоинство:** множественные связи между объектами.
- **Недостаток:** неизменность структуры после ввода данных.

Реляционная модель базы данных



- **Достоинство:** сравнительная простота инструментальных средств ее поддержки.
- **Недостаток:** зависимость скорости работы от размера базы данных.

Объектно-ориентированная модель БД

Достоинство: модель данных более близка сущностям реального мира. Типы данных определяются разработчиком и не ограничены набором предопределенных типов. Данные объекта и его методы составляют единое целое.

Недостаток: сложность реализации и сложность методологии.

Виды БД по технологии хранения и обработки данных

Централизованные БД	Распределенные БД
Расположение: один компьютер ***Назначение***: организация более простого и дешевого способа информационного обслуживания пользователей; ***Объемы данных***: небольшие ***Задачи***: несложные ***Надежность***: более высокая за счет организационной независимости	***Расположение:*** несколько компьютеров, объединенных в единую вычислительную систему с помощью вычислительных сетей; ***Назначение***: предоставление более гибких форм обслуживания множеству удаленных пользователей ***Объемы данных:*** значительные ***Задачи***: сложные ***Надежность*** обеспечивается за счет средств резервирования.

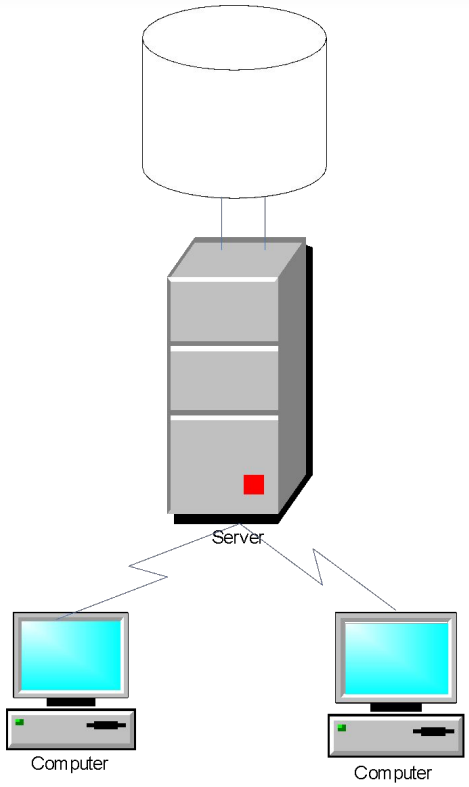
Условия централизации и децентрализации данных

- Данные централизуются, если:
  - данные непрерывно обновляются, а территориально разобщенные пользователи должны получать всякий раз последнее состояние данных;
  - поиск производится во всей совокупности данных;
  - над данными осуществляются операции со вторичными ключами.
- Данные могут быть децентрализованными, если они используются локально в точке их происхождения.
- При низкой скорости обновления допустимо хранение нескольких копий данных.

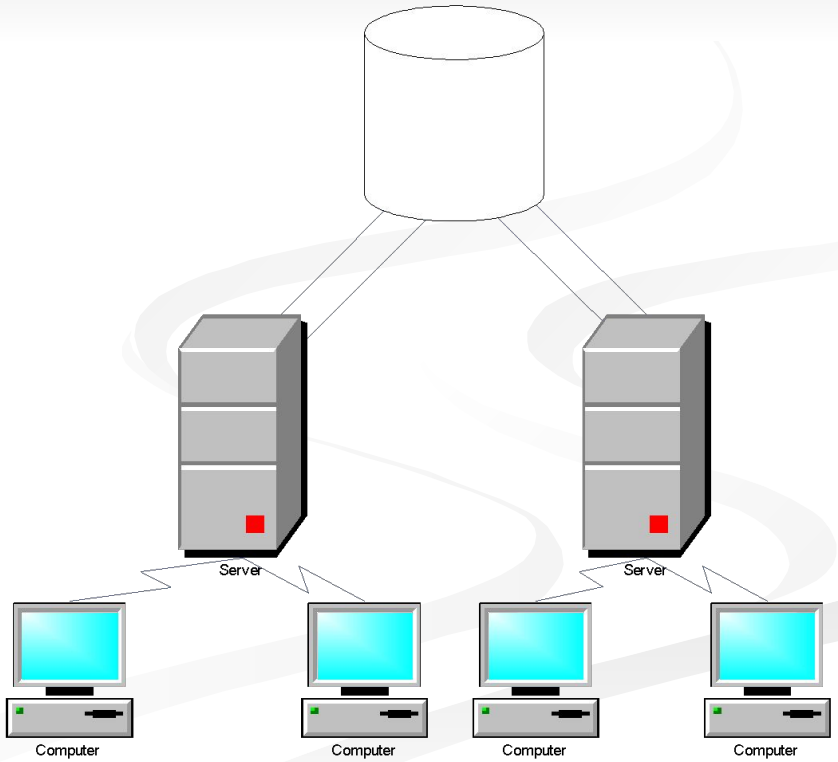
Классификация систем по способам распределения и обработки данных

Централизованные данные

Централизованные данные, централизованная обработка

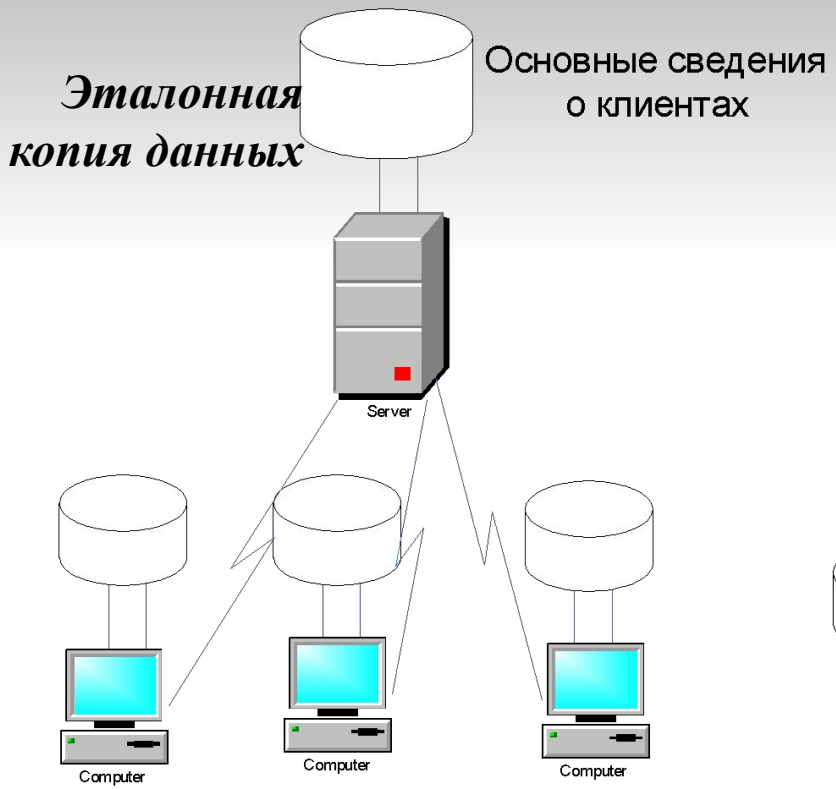


Централизованные данные, распределенная обработка

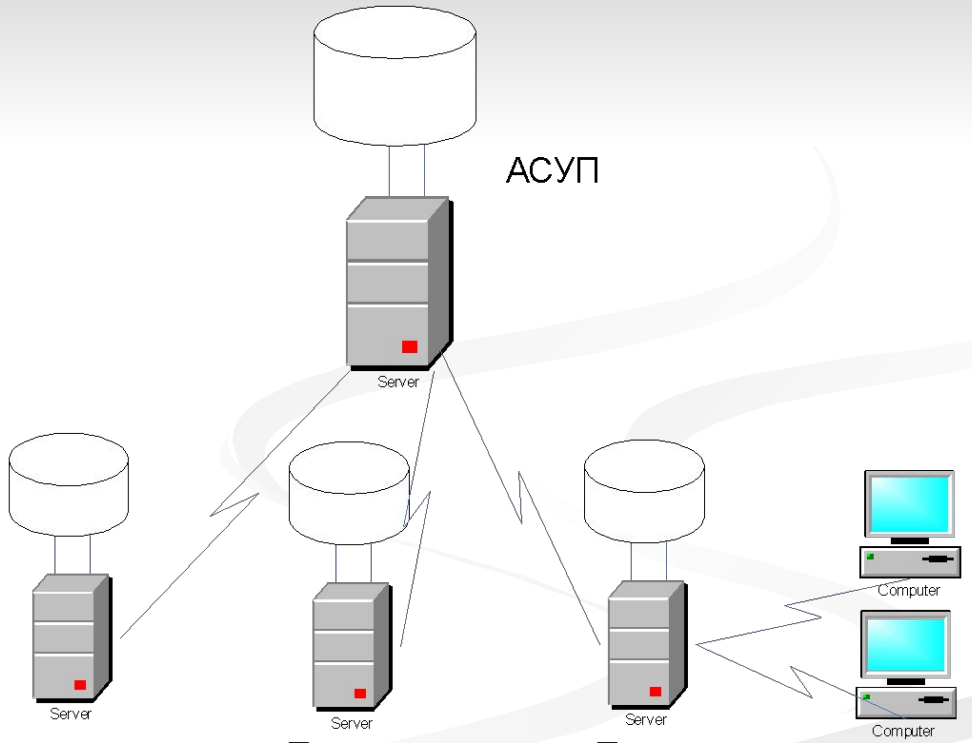


Иерархические данные

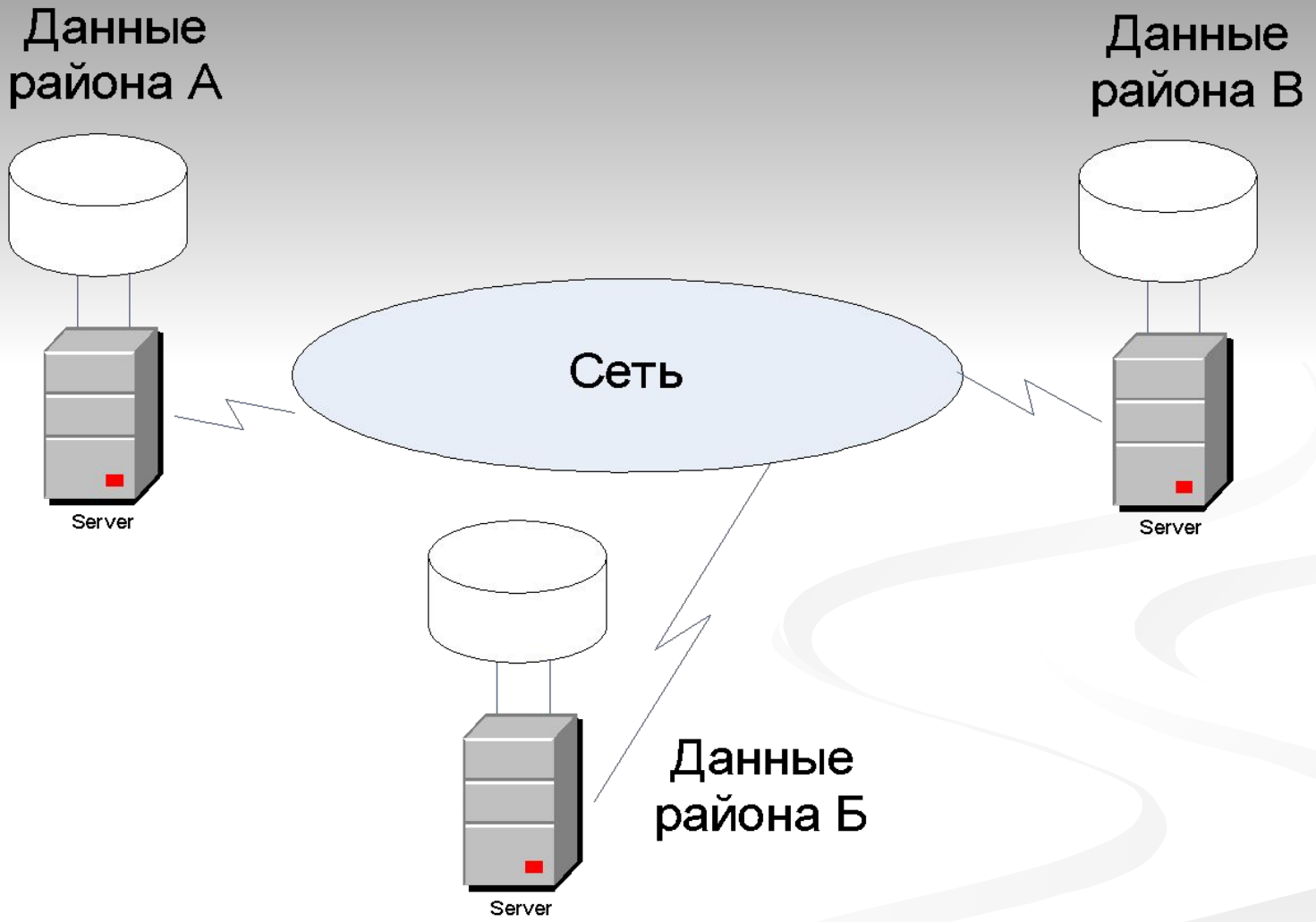
Зависимые данные



Независимые данные

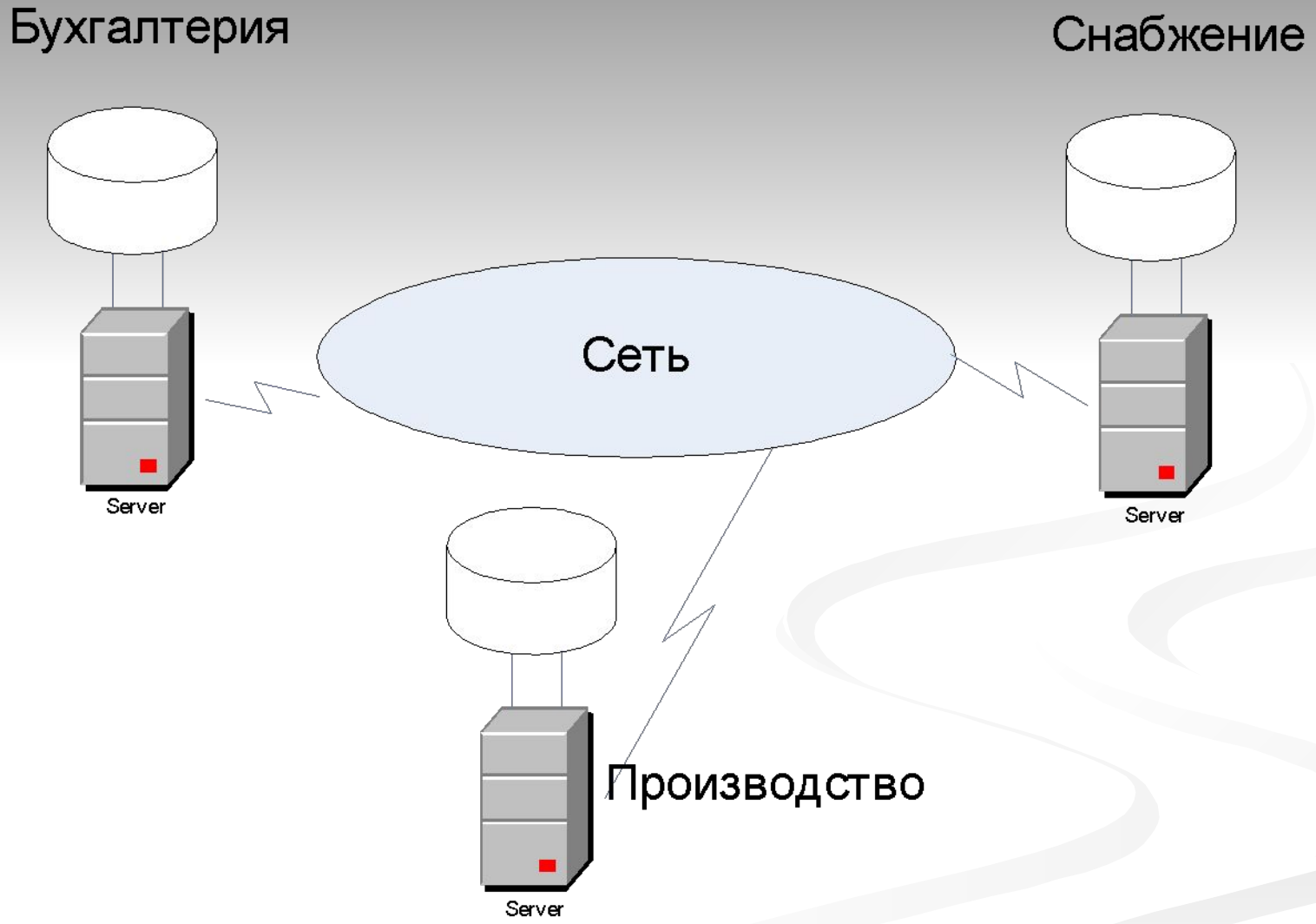


Расщепленные данные



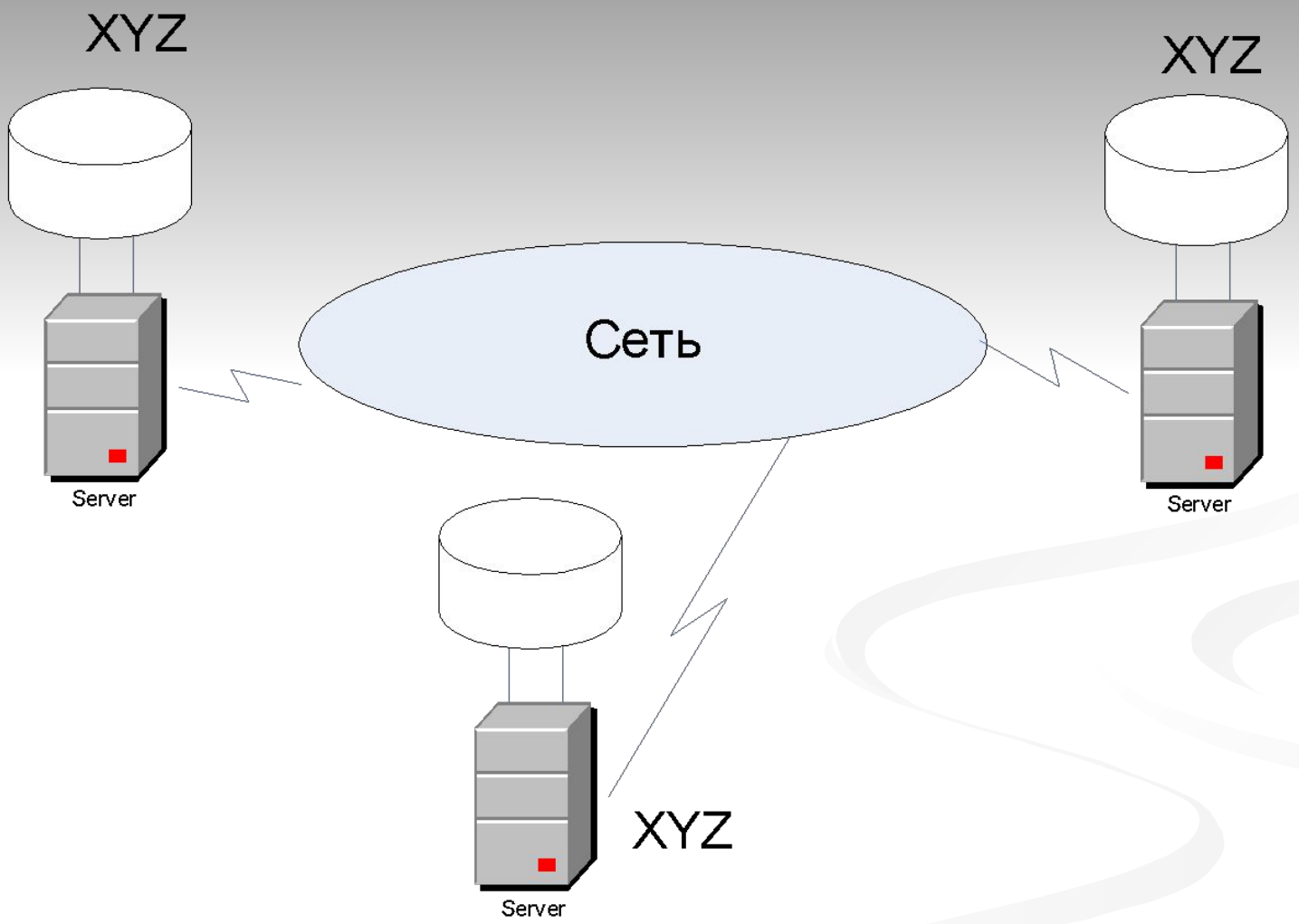
Структура данных и программы их обработки в подсистемах одни и те же. Содержание различно.

Разделенные данные



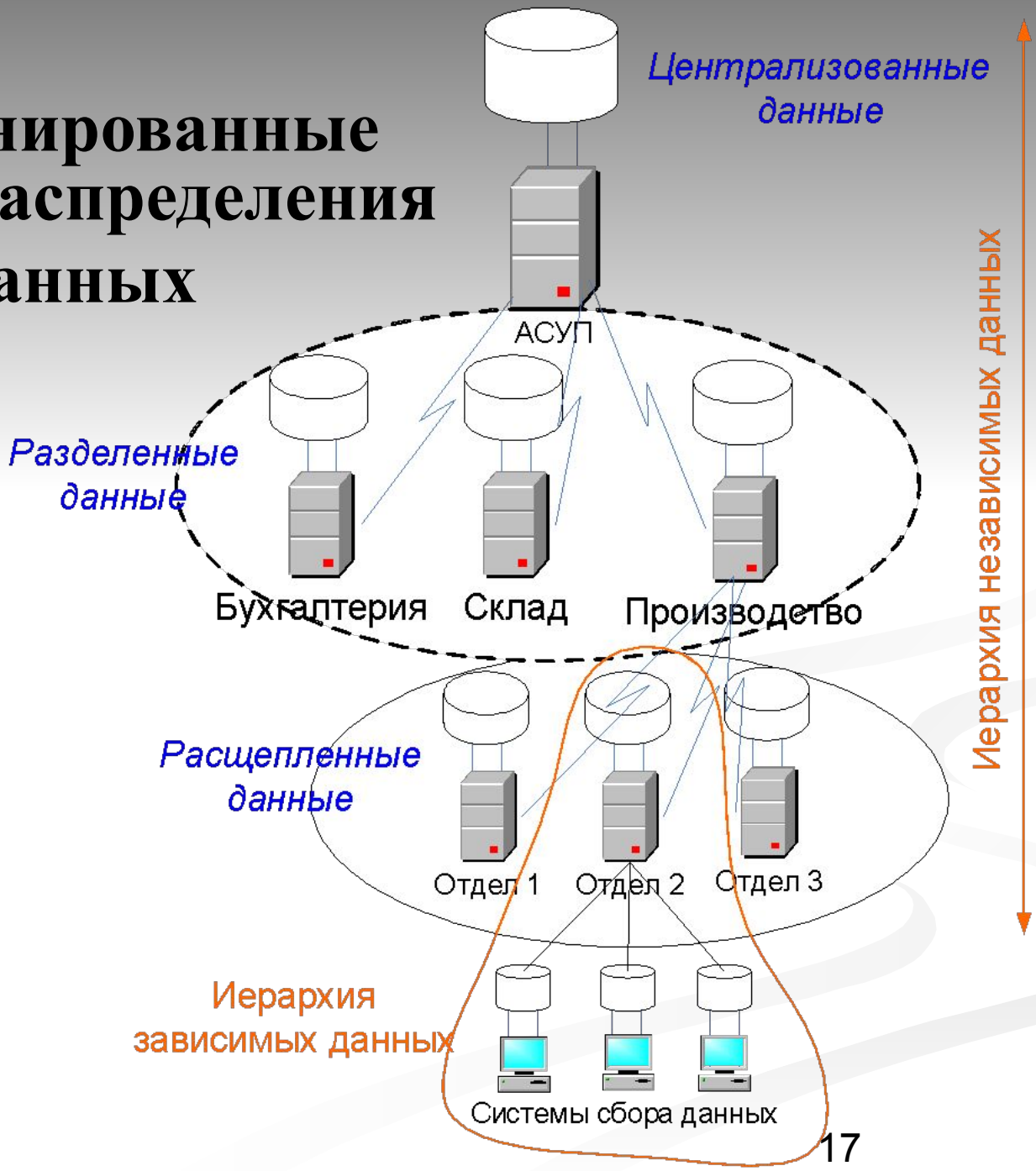
Структура данных, их содержание и программы обработки в подсистемах различны.

Реплицированные данные



Копии одних и тех же данных. Структура данных и программы обработки идентичны.

Комбинированные формы распределения данных



Организация ИО в виде банка данных

Банк данных – это автоматизированная система, представляющая совокупность информационных, программных, технических, языковых, организационно-методических средств и персонала, предназначенных для обеспечения централизованного накопления и коллективного многоцелевого использования данных.

Требования к банкам данных:

- интегрированность баз данных и целостность каждой из них;
- независимость и минимальная избыточность данных;
- способность к расширению.

Компоненты банка данных

- ***База данных;***
- ***Система управления базой данных;***
- ***Языковые средства*** – языки программирования, языки описания данных, языки запросов;
- ***Методические средства*** – инструкции и рекомендации по содержанию и функционированию банка данных, выбору СУБД;
- ***Технические средства*** – аппаратно-программный комплекс, на котором размещается БД и СУБД, удовлетворяющий по своим техническим характеристикам определенным требованиям;
- ***Персонал***
  - программисты,
  - инженеры по техническому обслуживанию аппаратно-программного комплекса,
  - администратор БД.

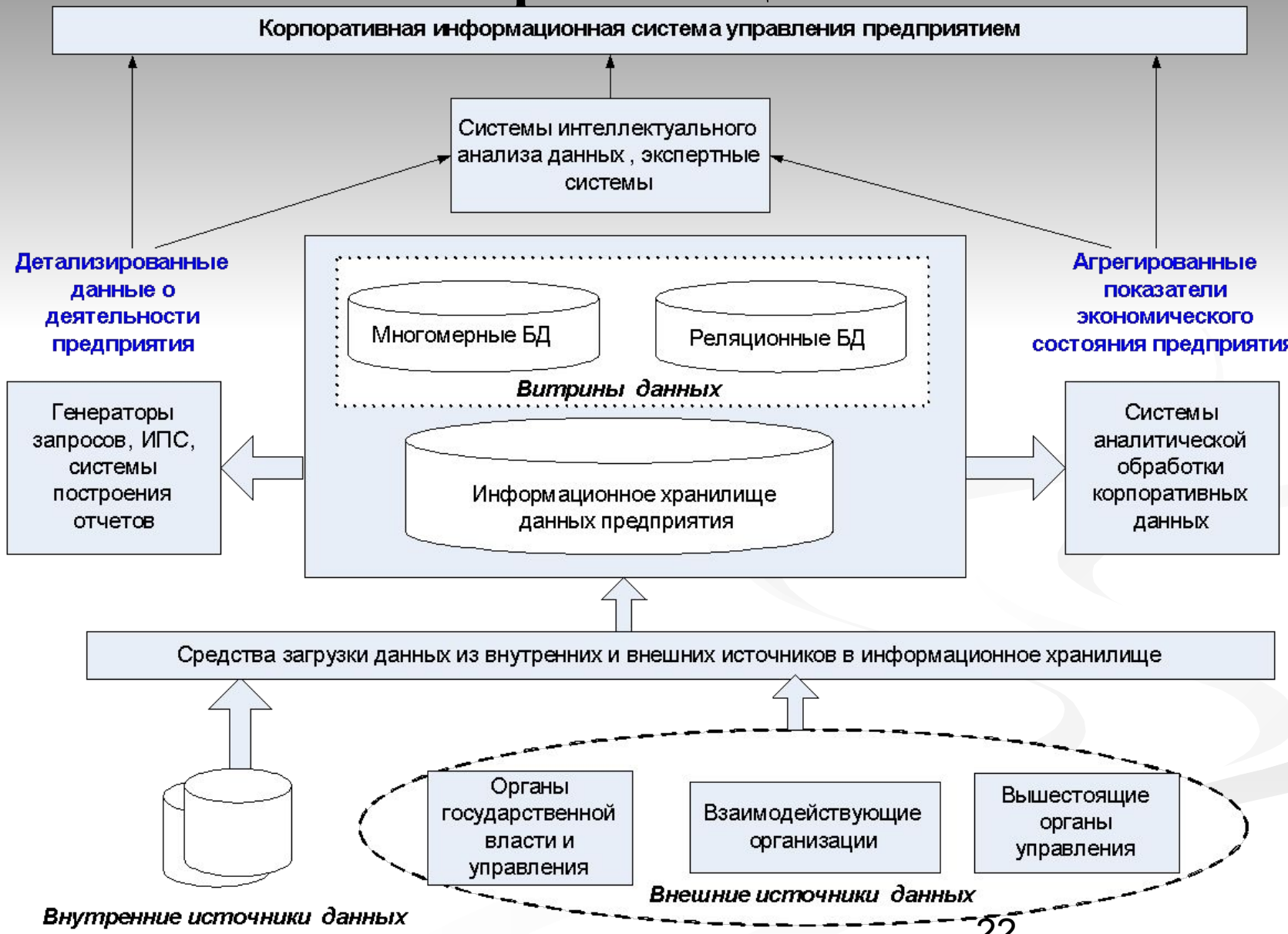
Концепция информационных хранилищ

- Информационное хранилище позволяет обеспечить:
  - хранение разнородных данных из различных источников в течение больших периодов времени;
  - быстрый доступ к данным и поиск релевантной запросу информации.

Причины появления информационных хранилищ

- Осознание руководством предприятий того, что в данных содержатся скрытые закономерности (знания), характеризующие процесс управления в целом, способные повысить его эффективность;
- снижение стоимости средств хранения информации, дающее возможность хранить данные, накопленные за длительные интервалы времени;
- снижение стоимости элементной базы сложных архитектур;
- переход от массового обслуживания к индивидуальному (учет разнообразных требований заказчика).

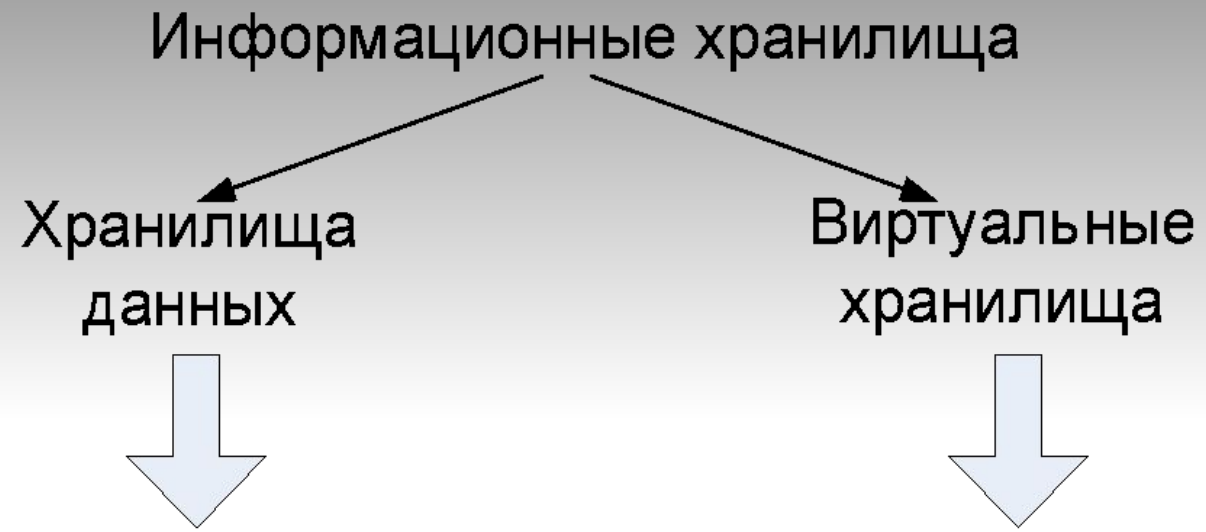
Концептуальная модель информационного хранилища



Проблемы интеграции данных

- Концепция информационных хранилищ подразумевает использование систем интеграции данных.
- Источники могут использовать различные модели данных и предоставлять различные интерфейсы для доступа к своим данным (реляционные, объектные или унаследованные СУБД).
- Данные источника могут быть неструктурированными (HTML файлы, текстовые файлы).
- Источники могут быть автономными.

Решение задачи интеграции данных



данные из различных источников поступают в хранилище, все запросы к системе интеграции обрабатываются с использованием этих данных.

данные хранятся в источниках, а запросы к системе интеграции транслируются в запросы или операции, понятные источнику. Данные, полученные в ответ на эти запросы к источникам, объединяются и предоставляются пользователю.

Хранилища данных

- ***Хранилище данных*** — это «предметно-ориентированная, интегрированная, содержащая исторические данные, неразрушаемая совокупность данных, предназначенная для поддержки принятия управленческих решений» (Уильям Инмон, 1992).
- ***Хранилище данных (Content Repository)*** – программная подсистема ИС, сочетающая в себе функции системы управления версиями, поисковой машины и СУБД.
- ***Хранилище данных (Data Warehouse)*** – очень большая предметно-ориентированная корпоративная база данных, специально разработанная и предназначенная для подготовки отчетов, анализа бизнес-процессов с целью поддержки принятия решений в организации.
- ***Хранилище данных*** – это автоматизированная информационно-технологическая система организации, которая собирает данные из существующих баз и внешних источников, формирует, хранит и эксплуатирует информацию в виде наборов данных.

Структура хранилища данных

Концепция хранилищ данных

Цель Хранилища Данных – подготовка данных к всестороннему анализу.

В основе концепции хранилища данных лежат две основные идеи:

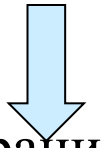
1. ***Интеграция*** ранее разъединенных детализированных данных в едином хранилище, их согласование и предварительная обработка.
2. ***Разделение*** хранящихся данных по их назначению – для операционной обработки и для использования в задачах анализа.

Концепция хранилищ данных

Цель Хранилища Данных – подготовка данных к всестороннему анализу.

В основе концепции хранилища данных лежат две основные идеи:

1. ***Интеграция*** ранее разъединенных детализированных данных в едином хранилище, их согласование и предварительная обработка.
2. ***Разделение*** хранящихся данных по их назначению – для операционной обработки и для использования в задачах анализа.



Процесс обработки данных в хранилище физически разделяется на два этапа.

1. Обработка транзакций в реальном времени (***OLTP – On-line Transaction Processing***), в результате чего в базах данных накапливается первичная информация о функционировании предприятия.
2. Аналитическая обработка данных в реальном времени (***OLAP – On-line Analytical Processing***).

Транзакционные и аналитические системы

При обработке корпоративной информации традиционным является разделение существующих задач на два класса:

- задачи оперативной обработки данных;
- задачи аналитической обработки данных.

Транзакционные системы ориентированы на операционную, или транзакционную обработку данных (автоматизированные информационные системы, осуществляющие учет и хранение оперативной информации по бизнес-процессам предприятия);

Аналитические системы ориентированы на анализ данных (системы поддержки принятия решений DSS - Decision Support System).

Признак	Транзакционная система	Аналитическая система
Цель	Учет, хранение и оперативная обработка непрерывно поступающих данных.	Получение и хранение обобщенных данных об объекте управления, предоставление информации для принятия решений.
Вид данных	Детализированные данные	Обобщенные данные
Частота обновления данных	Непрерывно, небольшими порциями	Редко
Представление результатов работы	Составление определенного набора отчетных форм	Получение большого числа разнообразных отчетов в удобном для понимания виде

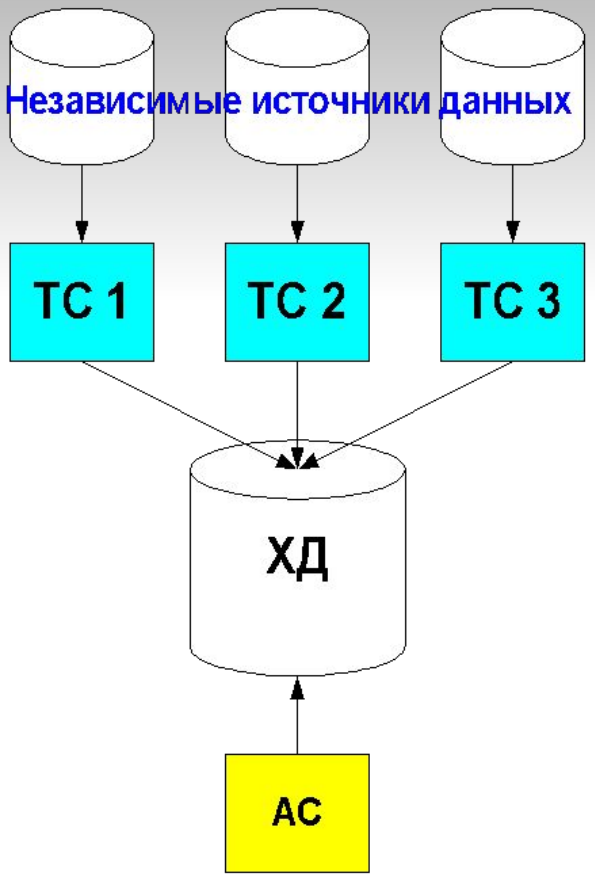
Характер использования системы

Транзакционная система	Аналитическая система
Автоматизация бизнес-процессов на уровне цехов, отделов, бюро.	▪ Получение на основе хранящихся данных показателей, определяющих закономерности развития предприятия и эффективность его работы. ▪ Предоставление средств и инструментов для обработки показателей с использованием различных методик анализа. ▪ Взаимодействие с различными программными пакетами, осуществляющими специализированную обработку данных (статистическими методами, с помощью нейронных сетей или нечеткой логики).

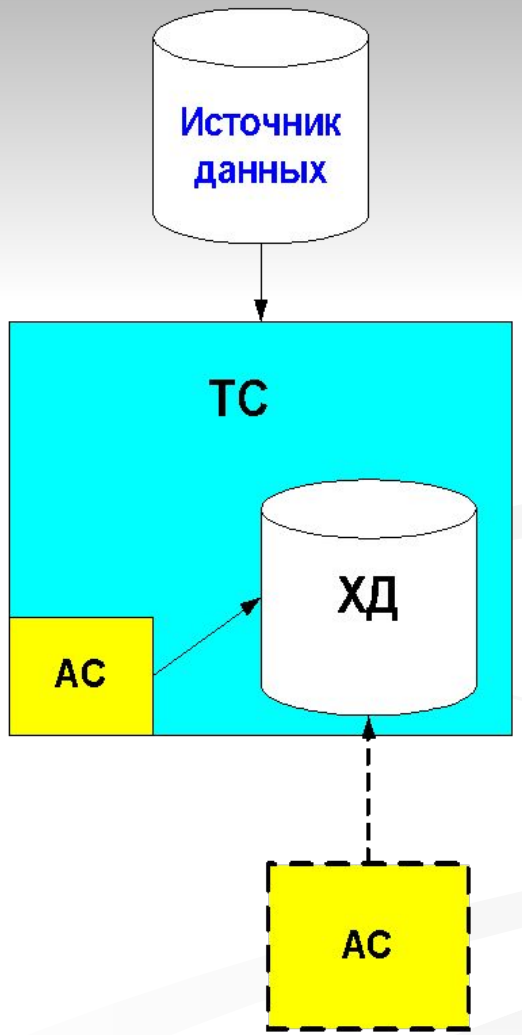
Взаимное сочетание транзакционной, аналитической систем и хранилища данных зависит от специфики деятельности организации, количества и характера информации.

Варианты использования ХД

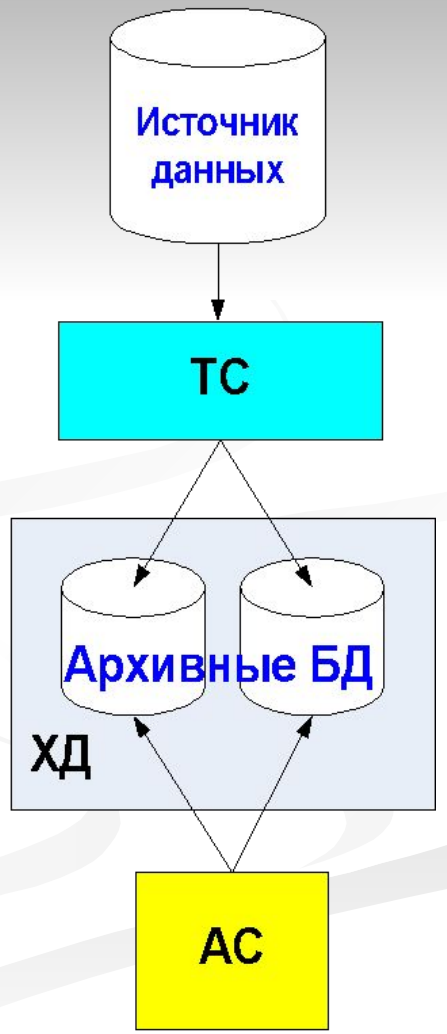
Вариант 1



Вариант 2



Вариант 3



АС – аналитическая система
ТС – транзакционная система
ХД – хранилище данных

Свойства данных

- ***Предметная ориентированность*** – все собираемые данные имеют отношение к определенной предметной области;
- ***Интегрированность*** – все данные взаимно согласованы и хранятся в едином Хранилище;
- ***Неизменяемость и целостность*** – исходные данные после переноса их в Хранилище, остаются неизменными и используются только в режиме чтения;
- ***Поддержка хронологии*** – данные хронологически структурированы и отражают историю за достаточный для выполнения задач анализа и прогноза период времени;
- ***Единство представления и удобство использования форм***.

Категории данных

Операции над данными

1. Сбор данных (пополнение хранилища данных)
   - очистка – устранение ненужной информации;
   - агрегирование – вычисление сумм, средних;
   - трансформация – преобразование типов данных, реорганизация структур хранения;
   - объединение из внешних и внутренних источников – приведение к единым форматам;
   - синхронизация – соответствие одному моменту времени.
2. Поддержка целостности и непротиворечивости данных
   - использование репозитория (словаря-справочника)
   - проверка данных на соответствие их структуре и назначению
3. Организация доступа к данным

Требования к хранилищам данных

- ***Высокая скорость загрузки данных.***
  - производительность процесса загрузки не должна накладывать ограничения на размер хранилища
- ***Обеспечение полнофункциональной технологии загрузки***
  - преобразование данных
  - фильтрация данных
  - переформатирование данных
  - проверка целостности данных
  - организация физического хранения данных
  - индексирование данных
  - обновление метаданных
- ***Высокое качество хранилища данных***
  - Мера качества хранилища – объективность исходных данных и степень разнообразия возможных запросов
- ***Поддержка различных видов данных***

Требования к хранилищам данных

- ***Высокая скорость обработки запросов***
  - зависит от сложности запроса, а не от объема хранилища
- ***Масштабируемость***.
  - поддержка СУБД параллельной обработки запросов
  - сохранение работоспособности в случае локальных аварий
  - обслуживание любого числа пользователей без потери производительности
- ***Широкие возможности администрирования***
  - контроль за приближением к ресурсным ограничениям
  - анализ затрат ресурсов
  - установка приоритетов для различных категорий пользователей и операций
  - осуществление настройки системы на максимальную производительность.

Витрины (киоски) данных

Витрина данных *(Data Mart)* – это тематическая база данных, содержащая информацию, относящуюся к отдельным аспектам деятельности организации.

Витрина данных является частью хранилища данных, специфицированной для использования конкретным подразделением или определенной группой пользователей.

Преимущества витрин данных	Недостатки витрин данных
1. Простота и невысокая стоимость реализации 2. Экономия технических ресурсов 3. Более высокий уровень безопасности данных 4. Высокая производительность	1. Дублирование данных 2. Необходимость синхронизации данных 3. Трудности расширения и объединения витрин 4. Ограниченность использования

Многоуровневое решение ХД

Виртуальные хранилища

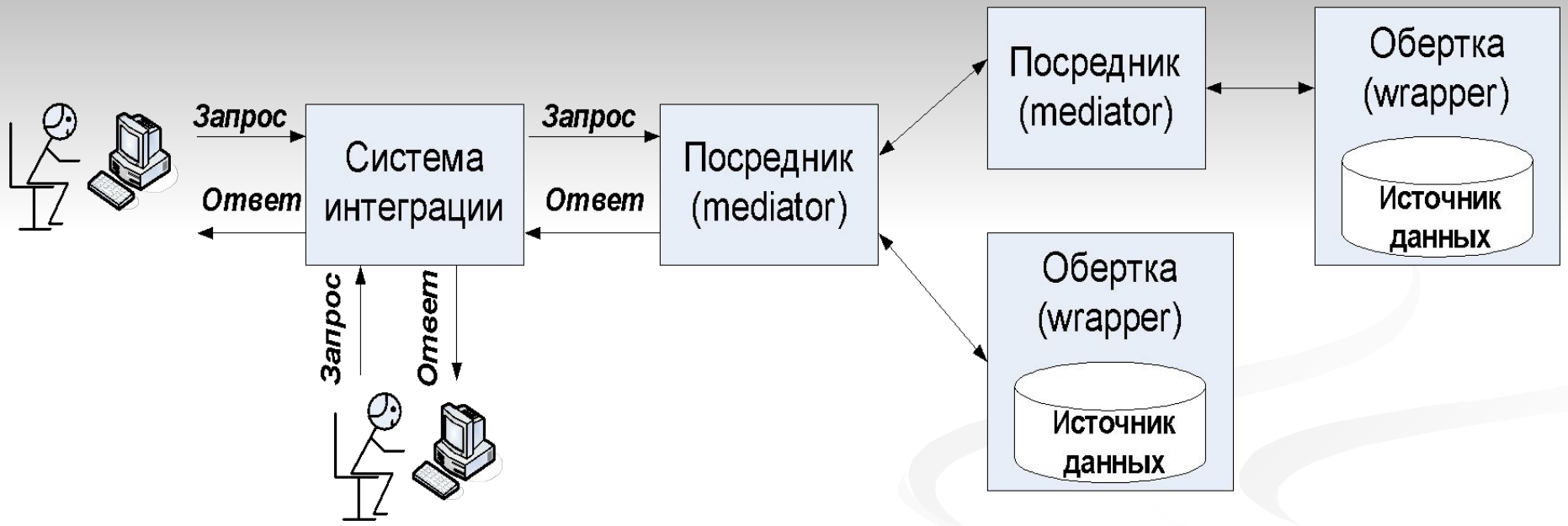
При использовании виртуальных хранилищ данные хранятся в удаленных источниках. Запрос к источнику транслируется через систему интеграции.

Достоинства	Недостатки
▪Всегда обновленные («свежие») данные ▪Простота и малая стоимость реализации ▪Единая платформа с источником информации ▪Отсутствие сетевых соединений между источником информации и хранилищем данных.	▪Сложность оптимизации запросов ▪Дополнительные расходы на конвертацию данных во время выполнения запроса ▪Более низкая производительность ▪Сложность интеграции данных с другими источниками ▪Отсутствие истории чистоты данных ▪Зависимость от доступности и структуры основной базы данных.

Логический уровень виртуального хранилища

- Логический уровень определяется выбором модели данных и языка запросов для этой модели.
- Модель используется для представления данных, извлекаемых из всех источников.
- Модель данных должна обеспечить прозрачность доступа к внешним источникам.
- Пользователь получает возможность унифицированного доступа ко всем интегрируемым данным, т.е. видит внешние данные как локальные в выбранной модели и не заботится об управлении доступом к источнику.

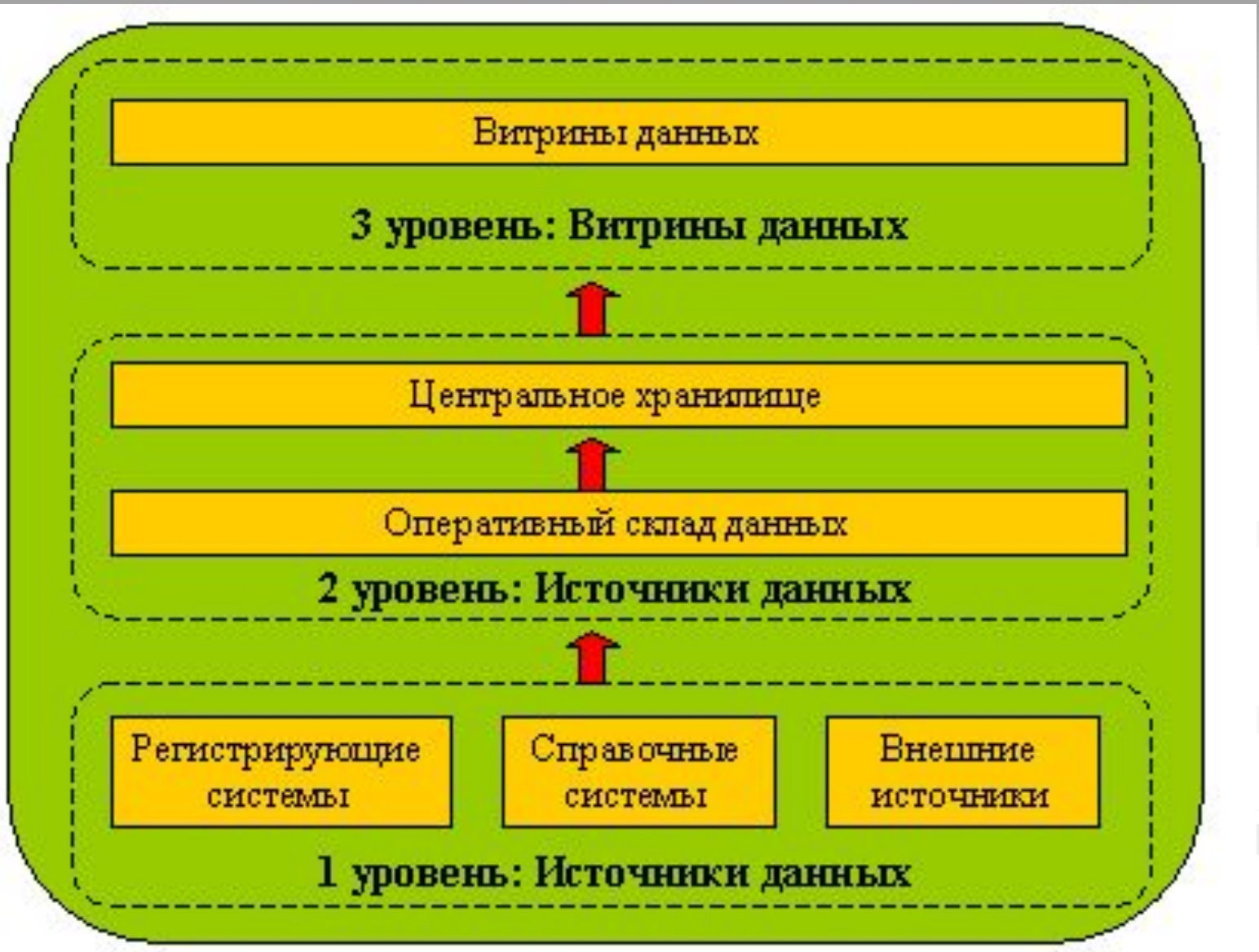
Физический уровень виртуального хранилища



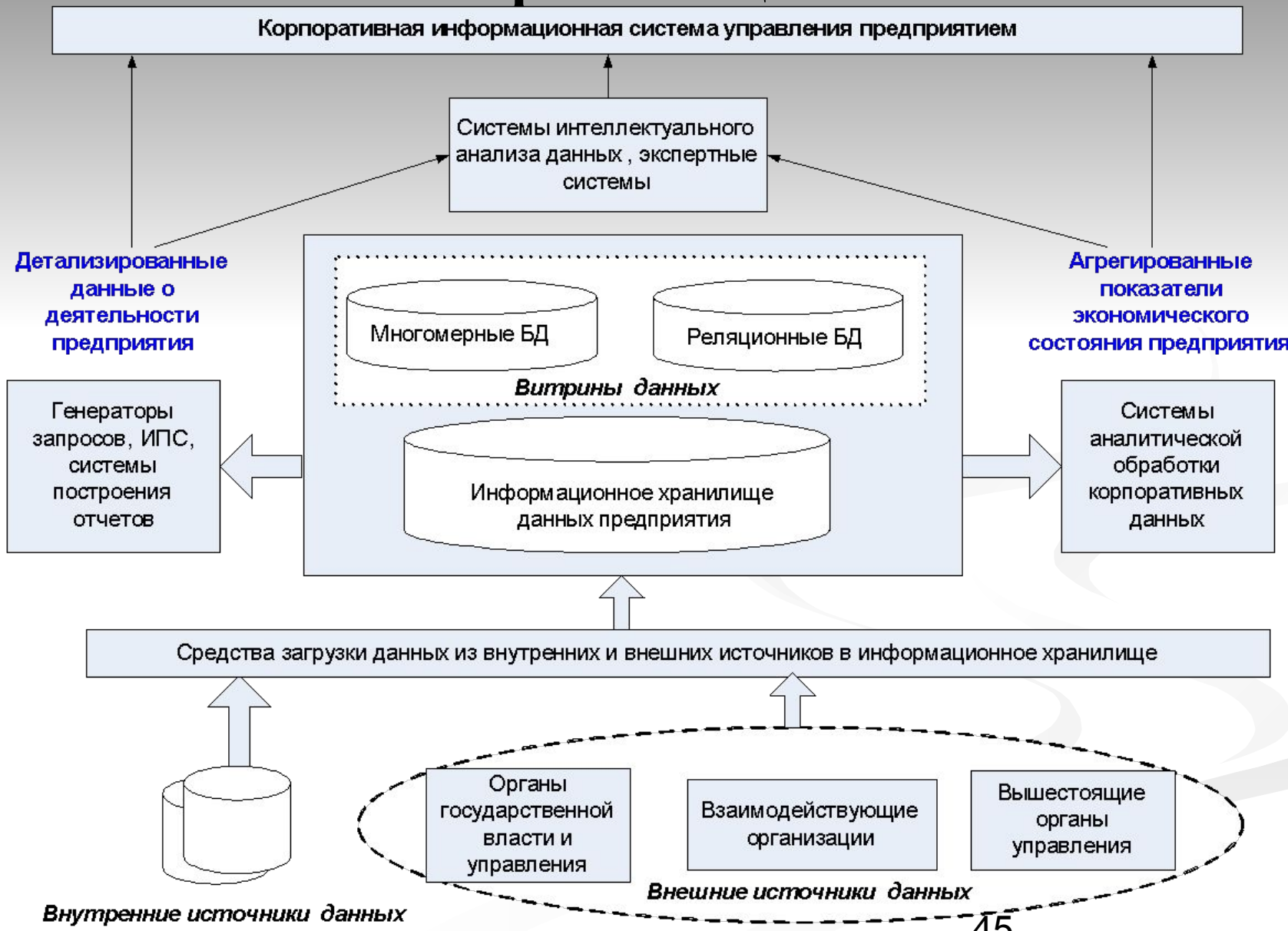
Обертка (wrapper) используется для хранения информации о внешнем источнике и организации к нему доступа.

Посредник (mediator) осуществляет интеграцию данных из различных источников

Корпоративное хранилище данных



Концептуальная модель информационного хранилища



Data Mining – это процесс обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах деятельности.

Примеры формулировок задач при использовании OLAP и Data Mining

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Какие факторы чаще всего определяют несчастные случаи?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов?	Какие характеристики отличают клиентов, которые собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Какие схемы покупок характерны для мошенничества с кредитными карточками?